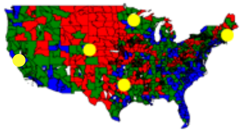
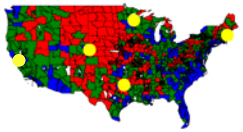


Adaptations, Extensions, and Practical Use of SMART Scores

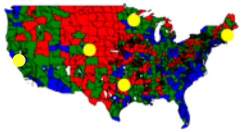
Ken Kleinman

Harvard Medical School and
Harvard Pilgrim Health Care



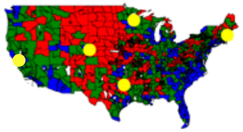
Outline

- What’s a SMART score?
- What’s so smart about them?
- Details of SMART score generation
- Can we make them smarter?
  - Easier models
  - Fewer fitting times
- Discussion



What’s a SMART score?

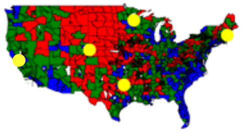
- SMART scores are a tool for evaluating surveillance data that comes in the form of reports from many regions
- SMART score: **SM**all **A**rea **R**egression and **T**esting score



What’s a SMART score?

Heuristically, we use regression to predict the count for each area on a given day. Then, we use statistical properties of the distribution to evaluate how unusual it is when the count is higher than the prediction.

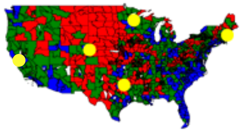
- Proposed in Kleinman et al., *Am J Epidemiol* 2004



What’s so smart about them?

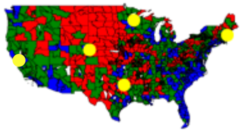
1. Relatively simple to generate
2. Can be implemented as two distinct steps, modeling and a look-up table.

The look-up table could be done on paper and is very simple on the computer.



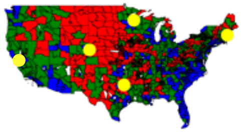
What’s so smart about them?

3. Lead nicely to ‘Recurrence Intervals.’ Like ‘100-year flood’ statements from the weather service, these are estimates of how often one should see results like the observed results, if nothing is happening. These were originally a byproduct of correcting for multiple comparisons.



Generating SMART scores

In the original formulation, I assumed that denominator, i.e. the number eligible to be cases, was available for every small area. (It was, in the original application.) This lead to a logistic regression formulation, essentially modeling the probability that any eligible was a case.

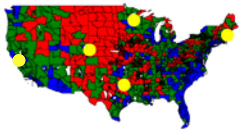


Generating SMART scores

- The model looks just like a logistic regression, with some additional subscripts and one more parameter:

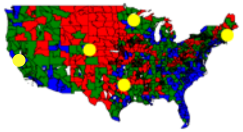
$$E(y_{it} \mid b_i) = n_{it} p_{it}; \qquad \mathrm{logit}(p_{it}) = x_{it}\beta + b_i$$

- where *i* is an area with counts on days *t*, y_{it} is the number of visits, n_{it} is the number of insured, and b_i is a random effect: b_i ~ N



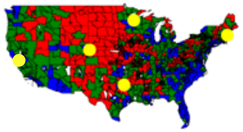
Generating SMART scores

- To use the model, we invert the estimated logit for each tract to get an estimated p_{it} for each area *i* on each day *t*
- Then we calculate the probability of seeing as many cases as we saw, or more, based on the binomial distribution
- The recurrence interval is a function of this p-value: $(p * 365 * \#\text{tracts})^{-1}$



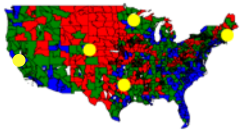
Generating SMART scores

- The random effect b_i models the unique features of each area: is there a little community of hypochondriacs somewhere? Are there more elderly?
- The estimated b_i (a.k.a. shrinkage or empircal Bayes estimators) are the odds of a case in area *i* relative to the average area.



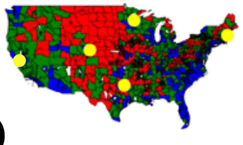
Example: Ambulatory Care

- Estimate effects of 11 months, 6 days of week, holiday, day after, time
- For Resp. Illness: Odds highest in winter months, lowest in summer
- Odds by day highest Mondays, lowest on weekends
- OR for holidays less than 1



Not so smart?

- Requires those denominators– can we do without them?
- Uses random effects– could we be simpler?
- What is the cost of making look-up tables vs. everyday fitting?
- Ignores spatial proximity
- General vs. recent trends/not T-S



Can we make it smarter?

- Requires those denominators– can we do without them?

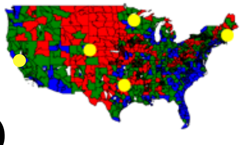
  →Poisson instead of Logistic?

- Uses random effects– could we be simpler?

  →Indicator variables?

- What is the cost of making look-up tables vs. everyday fitting?

  →Test in an example case

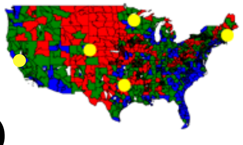


Can we make it smarter?

- Poisson model does not require denominators, though they are often used:

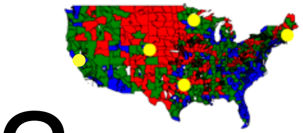
$$E(y_{it} \,|\, b_i) = \lambda_{it}; \qquad \log(\lambda_{it}) = x_{it}\beta + b_i$$

- Then get a p-value and recurrence interval as before, except that p-value is (of course) based on the Poisson distribution rather than binomial.



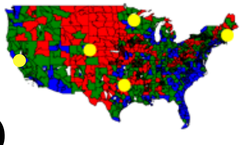
Can we make it smarter?

The Poisson is well-known as an approximation to the binomial. But one common rule of thumb is to have $N> 100$, $p< .01$ before using the approximation. Unclear how this pans out in this regression context, with N and p varying.



Can we make it smarter?

In theory, the impact of random effects for areas vs. indicator variables (fixed effects) for areas will be maximized when there are imbalances between the number of observations (days * denominator) across the areas. Here, we have exactly equal numbers of days, but unequal denominators. What will be the impact of indicators?



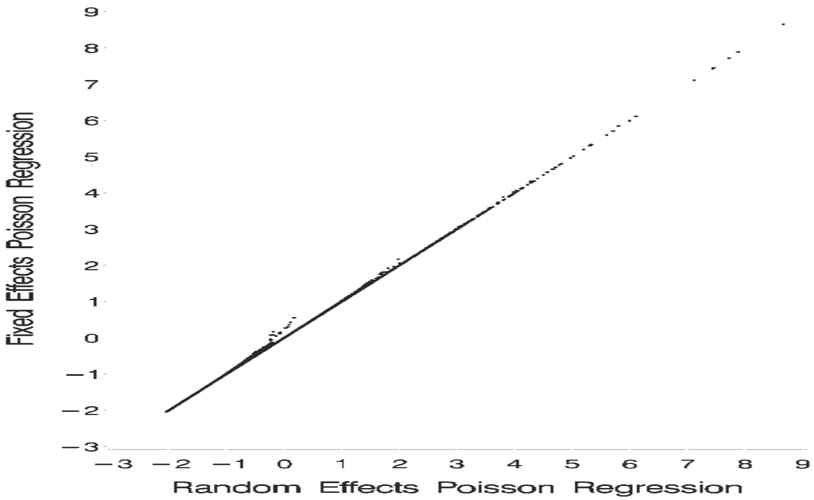
Can we make it smarter?

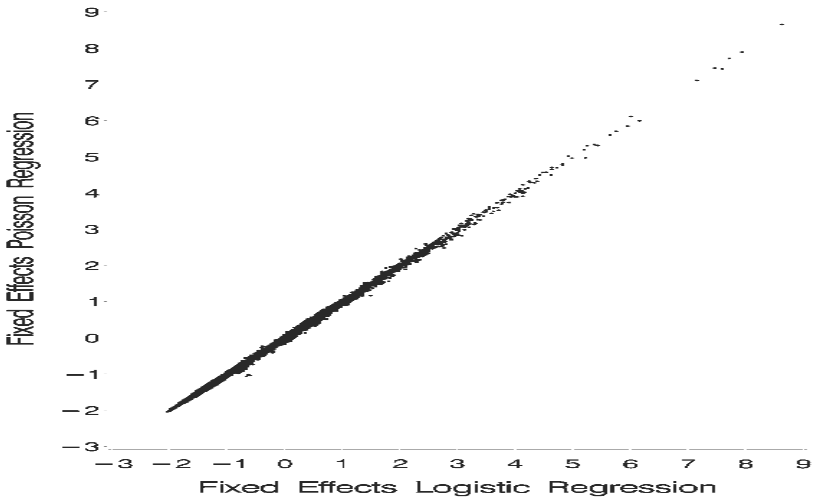
I will compare four (monthly) models:

1. Random Effects Logistic Regression
2. Random Effects Poisson Regression
3. Fixed Effects Logistic Regression
4. Fixed Effects Poisson Regression

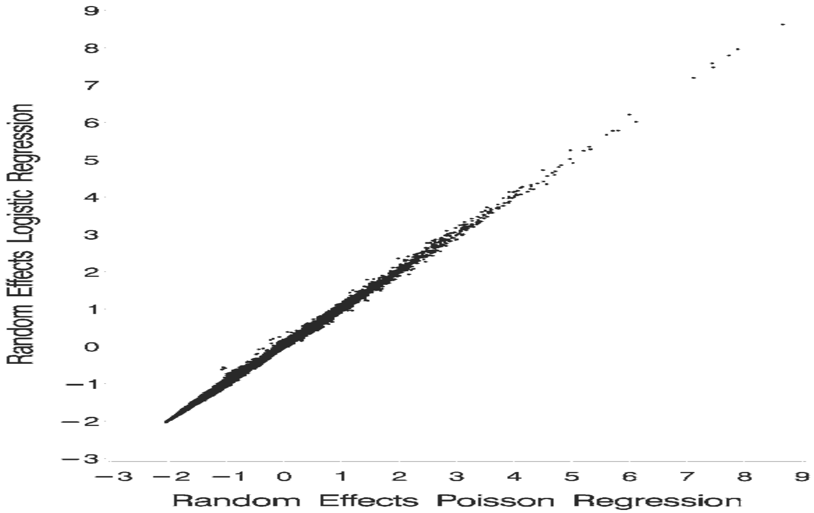
Each was used to generate RI for 1999 in 565 census tracts around Boston, based on respiratory complaints.

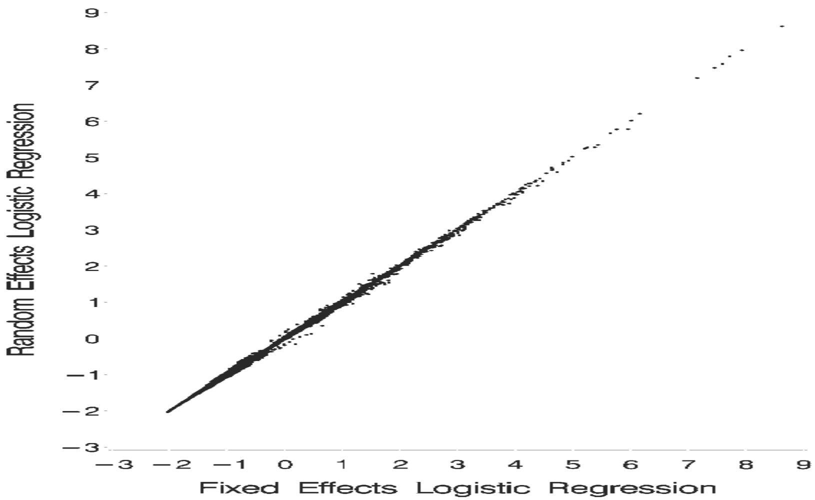
Results: log RI, continuous

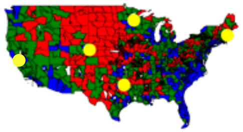




All: Correlation > 0.998

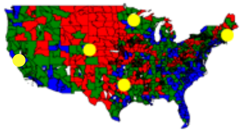






Results: by category

REPR RELR	0-180 days	181-365 days	366-1825	1826+ days
0-180 days	204806	3	0	0
181-365 days	18	77	8	0
366-1825	0	11	106	5
1826+ days	0	0	5	91

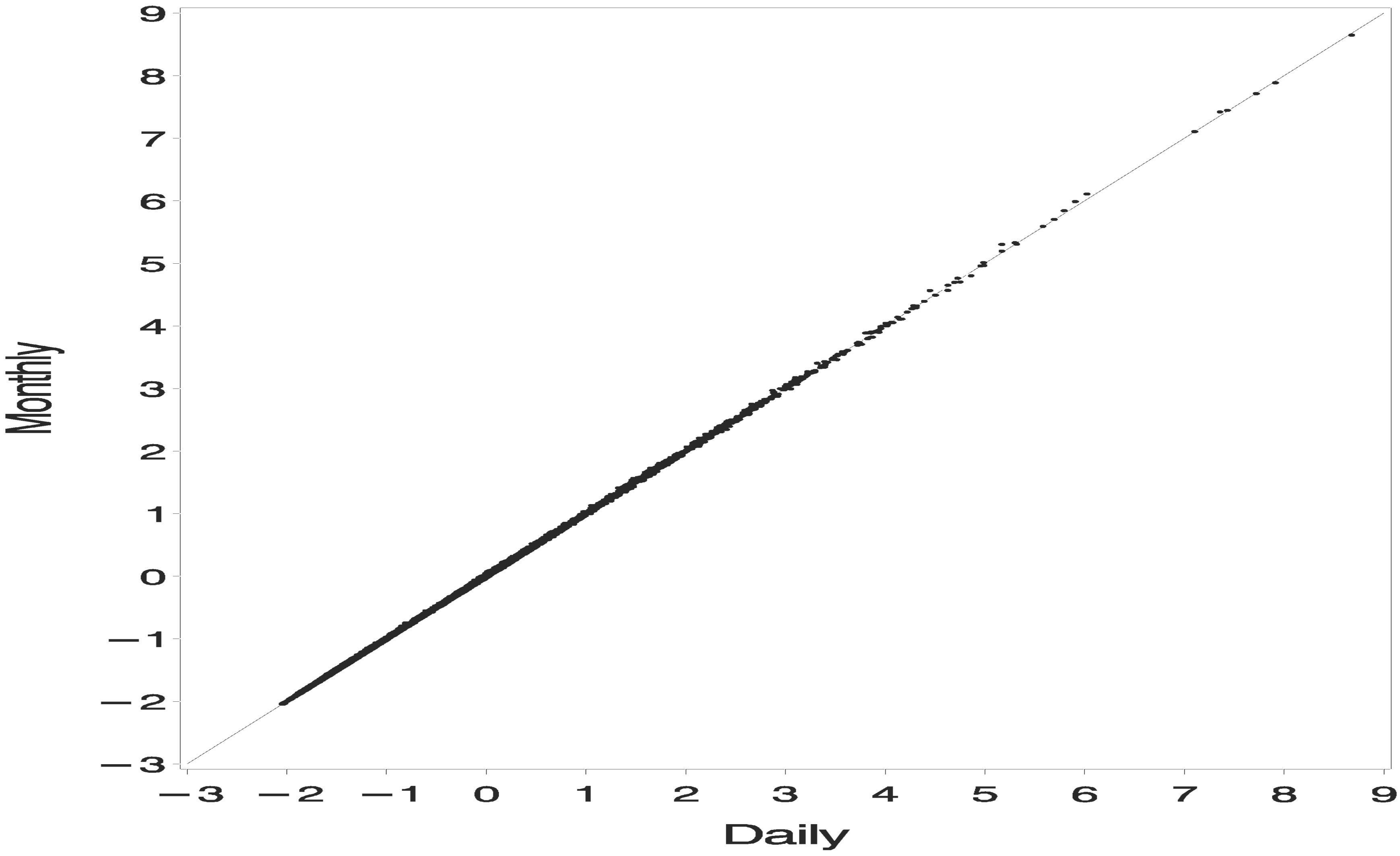


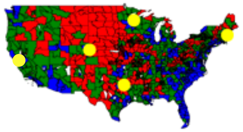
Can we make it smarter

If we can fit the models monthly, we can do it centrally, under supervision, without depending on successful automation.

To check the potential cost of this, I fit the fixed effects Poisson model daily, and compared the results for 1999 as before.

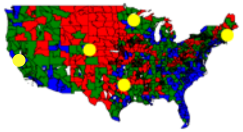
Results: log RI, continuous





Results: by category

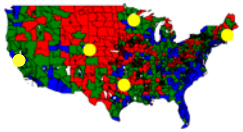
Daily Monthly	0-180 days	181-365 days	366-1825	1826+ days
0-180 days	204823	2	0	0
181-365 days	6	85	0	0
366-1825	0	1	115	2
1826+ days	0	0	3	93



Discussion

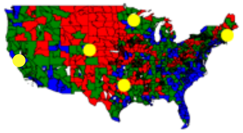
All four models generate very similar results based on monthly modeling. In this data, no need to require denominators or to use random effects models.

The daily models and monthly models result in nearly the same conclusions. We can use monthly modeling in this data.



Discussion

Some nice features derive from the property that the sum of two Poisson variates is Poisson with mean = sum of the means. For example we can make SMART scores that are based on the sum of cases across several days and/or several areas simply by summing the expected value on each day and/or in each area.



Postscript

A full discussion of this material, with more numeric results, is included as ‘Generalized linear models and generalized linear mixed models for small-area surveillance’ in

Spatial and Syndromic Surveillance for Public Health, Lawson and Kleinman, Eds., Wiley, 2005.